Chapter 5 Elements of Calculus

An Introduction to Optimization

Spring, 2014

Wei-Ta Chu

Sequences and Limits

- A sequence of real numbers can be viewed as a set of numbers $\{x_1, x_2, ..., x_k, ...\}$, which is often also denoted as $\{x_k\}$ or $\{x_k\}_{k=1}^{\infty}$
- A sequence $\{x_k\}$ is ***increasing*** if $x_1 < x_2 < \cdots < x_k < \cdots$. If $x_k \leq x_{k+1}$ then we say that the sequence is ***nondecreasing***. Similarly, we can define ***decreasing*** and ***nonincreasing*** sequences. Nonincreasing and nondecreasing sequences are called ***monotone sequences***.
- A number $x^* \in R$ is called the ***limit*** of the sequence $\{x_k\}$ if for any positive ϵ there is a number *K* (which may depend on ϵ) such that for all $k > K$, $|x_k - x^*| < \epsilon$. In this case, we write

  $$x^* = \lim_{k\to\infty} x_k \quad \text{or} \quad x_k \to x^*$$

- A sequence that has a limit is called a ***convergent sequence***.

Sequences and Limits

- A sequence in R^n is a function whose domain is the set of natural numbers 1, 2, …, k, … and whose range is contained in R^n. We use the notation $\{\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, ...\}$ or $\{\boldsymbol{x}^{(k)}\}$ for sequences in R^n.
- For limits of sequences in R^n, we need to replace absolution values with vector norms. In other words, $\boldsymbol{x}^*$ is the limit of $\{\boldsymbol{x}^{(k)}\}$ if for any positive ϵ there is a number K such that $k > K$, $\|\boldsymbol{x}^{(k)} - \boldsymbol{x}^*\| < \epsilon$.
- If a sequence $\{\boldsymbol{x}^{(k)}\}$ is convergent, we write $\boldsymbol{x}^* = \lim_{k\to\infty} \boldsymbol{x}^{(k)}$ or $\boldsymbol{x}^{(k)} \to \boldsymbol{x}^*$

Sequences and Limits

- Theorem 5.1: A convergent sequence has only one limit.
- A sequence $\{\boldsymbol{x}^{(k)}\}$ in R^n is ***bounded*** if there exists a number $B \geq 0$ such that $\|\boldsymbol{x}^{(k)}\| \leq B$ for all $k = 1, 2, ...$
- Theorem 5.2: Every convergent sequence is bounded.
- For a sequence $\{x_k\}$ in *R*, a number *B* is called an ***upper bound*** if $x_k \leq B$ for all $k = 1, 2, ...$ In this case, we say $\{x_k\}$ is ***bounded above***.
- A number *B* is called an ***lower bound*** if $x_k \geq B$ for all $k = 1, 2, ...$ In this case, we say $\{x_k\}$ is ***bounded below***.

Sequences and Limits

- Any sequence $\{x_k\}$ in *R* that has an upper bound has a ***least upper bound*** (also called the ***supremum***), which is the smallest number *B* that is an upper bound of $\{x_k\}$. Similarly, it has a ***greatest lower bound*** (also called ***infimum***).
- If *B* is the least upper bound of the sequence $\{x_k\}$, then $x_k \leq B$ for all k, and for any $\epsilon > 0$, there exists a number *K* such that $x_K > B - \epsilon$
- If *B* is the greatest lower bound of $\{x_k\}$, then $x_k \geq B$ for all k, and for any $\epsilon > 0$, there exists a number *K* such that $x_K < B + \epsilon$
- Theorem 5.3: Every monotone bounded sequence in *R* is convergent.

Sequences and Limits

- Given a sequence $\{\boldsymbol{x}^{(k)}\}$ and an increasing sequence of natural numbers $\{m_k\}$, the sequence
$$\{\boldsymbol{x}^{(m_k)}\} = \{\boldsymbol{x}^{(m_1)}, \boldsymbol{x}^{(m_2)}, ...\}$$
is called a ***subsequence*** of the sequence $\{\boldsymbol{x}^{(k)}\}$.
- Theorem 5.4: Consider a convergent sequence $\{\boldsymbol{x}^{(k)}\}$ with limit $\boldsymbol{x}^*$. Then any subsequence of $\{\boldsymbol{x}^{(k)}\}$ also converges to $\boldsymbol{x}^*$.
- It turns out that any bounded sequence contains a convergent subsequence (***Bolzano-Weierstrass Theorem***)

Sequences and Limits

- Consider a function $\boldsymbol{f} : R^n \to R^m$ and a point $\boldsymbol{x}_0 \in R^n$. Suppose that there exists $\boldsymbol{f}^*$ such that for any convergent sequence $\{\boldsymbol{x}^{(k)}\}$ with limit $\boldsymbol{x}_0$, we have

  $$\lim_{k\to\infty} \boldsymbol{f}(\boldsymbol{x}^{(k)}) = \boldsymbol{f}^*$$

  Then, we use the notation $\lim_{\boldsymbol{x}\to\boldsymbol{x}_0} \boldsymbol{f}(\boldsymbol{x})$ to represent $\boldsymbol{f}^*$
- It turns out that $\boldsymbol{f}$ is continuous at $\boldsymbol{x}_0$ if and only if for any convergent sequence $\{\boldsymbol{x}^{(k)}\}$ with limit $\boldsymbol{x}_0$, we have

  $$\lim_{k\to\infty} \boldsymbol{f}(\boldsymbol{x}^{(k)}) = \boldsymbol{f}\left(\lim_{k\to\infty} \boldsymbol{x}^{(k)}\right) = \boldsymbol{f}(\boldsymbol{x}_0)$$
- Therefore, using the notation introduced above, the function $\boldsymbol{f}$ is continuous at $\boldsymbol{x}_0$ if and only if

  $$\lim_{\boldsymbol{x}\to\boldsymbol{x}_0} \boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{f}(\boldsymbol{x}_0)$$

Sequences and Limits

- We say that a sequence $\{\boldsymbol{A}_k\}$ of $m \times n$ matrices converges to the $m \times n$ matrix $\boldsymbol{A}$ if $\lim_{k\to\infty}\|\boldsymbol{A} - \boldsymbol{A}_k\| = 0$.
- Lemma 5.1: Let $\boldsymbol{A} \in R^{n\times n}$. Then, $\lim_{k\to\infty} \boldsymbol{A}^k = \boldsymbol{O}$ if and only if the eigenvalues of $\boldsymbol{A}$ satisfy $|\lambda_i(\boldsymbol{A})| < 1, i = 1, ..., n$
- Lemma 5.2: The series of $n \times n$ matrices

$$\boldsymbol{I}_n + \boldsymbol{A} + \boldsymbol{A}^2 + \cdots + \boldsymbol{A}^k + \cdots$$

converges if and only if $\lim_{k\to\infty} \boldsymbol{A}^k = \boldsymbol{O}$. In this case the sum of the series equals $(\boldsymbol{I}_n - \boldsymbol{A})^{-1}$

Sequences and Limits

- A matrix-valued function $\boldsymbol{A} : R^r \rightarrow R^{n \times n}$ is continuous at a point $\boldsymbol{\xi}_0 \in R^r$ if $\lim_{\|\boldsymbol{\xi} - \boldsymbol{\xi}_0\|} \|\boldsymbol{A}(\boldsymbol{\xi}) - \boldsymbol{A}(\boldsymbol{\xi}_0)\| = 0$
- Lemma 5.3: Let $\boldsymbol{A} : R^r \rightarrow R^{n \times n}$ be an $n \times n$ matrix-valued function that is continuous at $\boldsymbol{\xi}_0$. If $\boldsymbol{A}(\boldsymbol{\xi}_0)^{-1}$exists, then $\boldsymbol{A}(\boldsymbol{\xi})^{-1}$ exists for $\boldsymbol{\xi}$ sufficiently close to $\boldsymbol{\xi}_0$ and $\boldsymbol{A}(\cdot)^{-1}$ is continuous at $\boldsymbol{\xi}_0$.

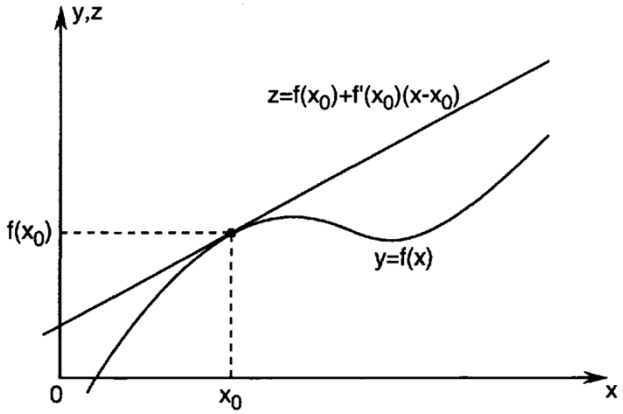
Differentiability

- Differential calculus is based on the idea of approximating an arbitrary function by an ***affine function***.
- A function $\mathcal{A} : R^n \to R^m$ is ***affine*** if there exists a ***linear function*** $\mathcal{L} : R^n \to R^m$ and a vector $\boldsymbol{y} \in R^m$ such that

$$\mathcal{A}(\boldsymbol{x}) = \mathcal{L}(\boldsymbol{x}) + \boldsymbol{y}$$

  for every $\boldsymbol{x} \in R^n$
- Consider a function $\boldsymbol{f} : R^n \to R^m$ and a point $\boldsymbol{x}_0 \in R^n$. We wish to find an affine function $\mathcal{A}$ that approximates $\boldsymbol{f}$ near the point $\boldsymbol{x}_0$
- First, it is natural to impose the condition

$$\mathcal{A}(\boldsymbol{x}_0) = \boldsymbol{f}(\boldsymbol{x}_0)$$



$$\mathcal{A}(\boldsymbol{x}_0) = \boldsymbol{f}(\boldsymbol{x}_0)$$

Differentiability

- Because $\mathcal{A}(\boldsymbol{x}) = \mathcal{L}(\boldsymbol{x}) + \boldsymbol{y}$, we obtain $\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{x}_0) - \mathcal{L}(\boldsymbol{x}_0)$
- By the linearity of $\mathcal{L}$,

$$\mathcal{L}(\boldsymbol{x}) + \boldsymbol{y} = \mathcal{L}(\boldsymbol{x}) - \mathcal{L}(\boldsymbol{x}_0) + \boldsymbol{f}(\boldsymbol{x}_0) = \mathcal{L}(\boldsymbol{x} - \boldsymbol{x}_0) + \boldsymbol{f}(\boldsymbol{x}_0)$$

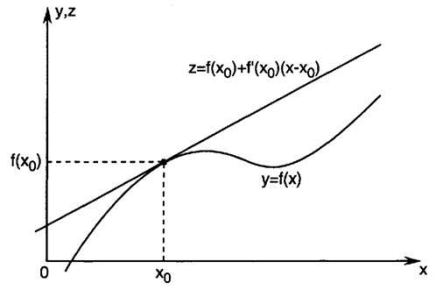
- Hence, we may write

$$\mathcal{A}(\boldsymbol{x}) = \mathcal{L}(\boldsymbol{x} - \boldsymbol{x}_0) + \boldsymbol{f}(\boldsymbol{x}_0)$$

- Next, we require that $\mathcal{A}(\boldsymbol{x})$ approaches $\boldsymbol{f}(\boldsymbol{x})$ faster than $\boldsymbol{x}$ approaches $\boldsymbol{x}_0$; that is,

$$\lim_{\boldsymbol{x} \to \boldsymbol{x}_0, \boldsymbol{x} \in \Omega} \frac{\|\boldsymbol{f}(\boldsymbol{x}) - \mathcal{A}(\boldsymbol{x})\|}{\|\boldsymbol{x} - \boldsymbol{x}_0\|} = 0$$

- The conditions ensure that $\mathcal{A}$ approximates $\boldsymbol{f}$ near $\boldsymbol{x}_0$ in the sense that the approximation error is “small” compared with the distance of the point from $\boldsymbol{x}_0$.



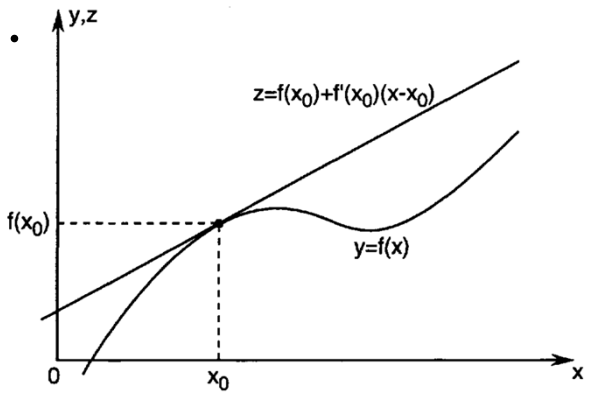
$$\mathcal{A}(\boldsymbol{x}) = \mathcal{L}(\boldsymbol{x} - \boldsymbol{x}_0) + \boldsymbol{f}(\boldsymbol{x}_0) \quad \lim_{\boldsymbol{x}\to\boldsymbol{x}_0, \boldsymbol{x}\in\Omega} \frac{\|\boldsymbol{f}(\boldsymbol{x}) - \mathcal{A}(\boldsymbol{x})\|}{\|\boldsymbol{x} - \boldsymbol{x}_0\|} = 0$$

Differentiability

- In summary, a function $\boldsymbol{f} : \Omega \to R^m, \Omega \subset R^n,$ is said to be ***differentiable*** at $\boldsymbol{x}_0 \in \Omega$ if there is an affine function that approximates $\boldsymbol{f}$ near $\boldsymbol{x}_0$; that is, there exists a linear function $\mathcal{L} : R^n \to R^m$ such that

$$\lim_{\boldsymbol{x}\to\boldsymbol{x}_0, \boldsymbol{x}\in\Omega} \frac{\|\boldsymbol{f}(\boldsymbol{x}) - (\mathcal{L}(\boldsymbol{x} - \boldsymbol{x}_0) + \boldsymbol{f}(\boldsymbol{x}_0))\|}{\|\boldsymbol{x} - \boldsymbol{x}_0\|} = 0$$

- The linear function $\mathcal{L}$ is determined uniquely by $\boldsymbol{f}$ and $\boldsymbol{x}_0$ and is called the ***derivative*** of $\boldsymbol{f}$ at $\boldsymbol{x}_0$.
- The function is said to be ***differentiable*** on Ω if $\boldsymbol{f}$ is differentiable at every point of its domain Ω .



Differentiability

- In R, an affine function has the form $ax + b$, with $a, b \in R$. Hence, a real-valued function $f(x)$ of a real variable x that is differentiable at x_0 can be approximated by a function

$$\mathcal{A}(x) = ax + b$$

- Because $f(x_0) = \mathcal{A}(x_0) = ax_0 + b$, we obtain

$$\mathcal{A}(x) = ax + b = a(x - x_0) + f(x_0)$$

- The linear part of $\mathcal{A}(x)$, denoted by $\mathcal{L}(x)$ earlier, is just ax. The norm of a real number is its absolute value, so by the definition of differentiability

$$\lim_{\boldsymbol{x} \to \boldsymbol{x}_0} \frac{|f(x) - (a(x - x_0) + f(x_0))|}{|x - x_0|} = 0$$

$$\Rightarrow \quad \lim_{\boldsymbol{x} \to \boldsymbol{x}_0} \frac{|f(x) - f(x_0)|}{|x - x_0|} = a$$

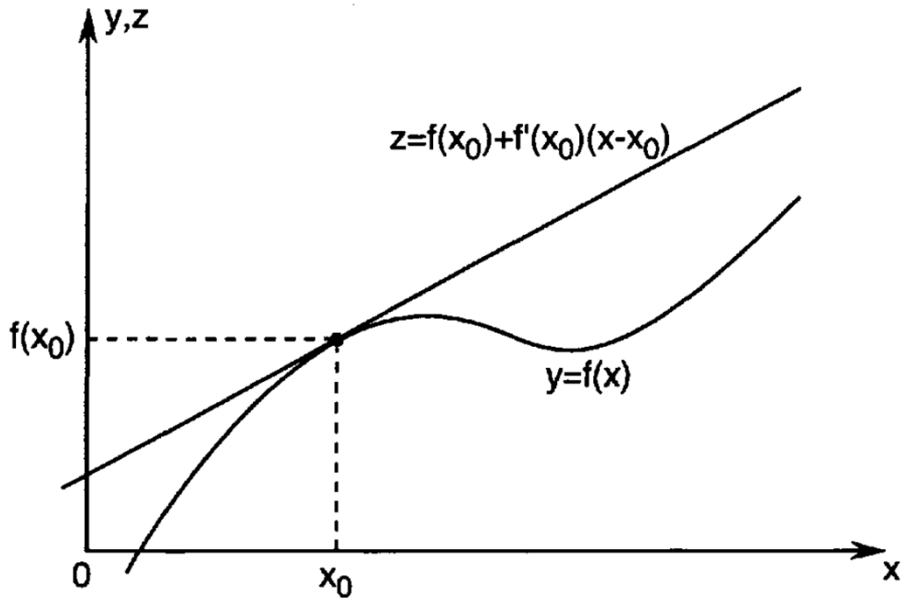
$$\lim_{x \to x_0} \frac{|f(x) - f(x_0)|}{|x - x_0|} = a \qquad \mathcal{A}(x) = ax + b = a(x - x_0) + f(x_0)$$

Differentiability

- The number a is commonly denoted $f'(x_0)$ and is called the derivative of f at x_0.
- The affine function $\mathcal{A}$ is therefore given by

$$\mathcal{A}(x) = f(x_0) + f'(x_0)(x - x_0)$$

- The affine function is tangent to f at x_0.



The Derivative Matrix

$$\lim_{\boldsymbol{x}\to\boldsymbol{x}_0,\boldsymbol{x}\in\Omega}\frac{\|\boldsymbol{f}(\boldsymbol{x})-(\mathcal{L}(\boldsymbol{x}-\boldsymbol{x}_0)+\boldsymbol{f}(\boldsymbol{x}_0))\|}{\|\boldsymbol{x}-\boldsymbol{x}_0\|}=0$$

- Any linear transformation from R^n to R^m, and in particular the derivative $\mathcal{L}$ of $\boldsymbol{f}: R^n \to R^m$, can be represented by an $m \times n$ matrix.
- To find the matrix $\boldsymbol{L}$ of the derivative $\mathcal{L}$, we use the natural basis $\{\boldsymbol{e}_1, \boldsymbol{e}_2, ..., \boldsymbol{e}_n\}$ for R^n. Consider the vectors

$$\boldsymbol{x}_j = \boldsymbol{x}_0 + t\boldsymbol{e}_j, j = 1, 2, .., n$$

By the definition of the derivative, we have

$$\lim_{t\to 0}\frac{\boldsymbol{f}(\boldsymbol{x}_j)-(t\boldsymbol{L}\boldsymbol{e}_j+\boldsymbol{f}(\boldsymbol{x}_0))}{t}=\boldsymbol{0} \qquad j=1,2,..,n$$

This means that

$$\lim_{t\to 0}\frac{\boldsymbol{f}(\boldsymbol{x}_j)-\boldsymbol{f}(\boldsymbol{x}_0)}{t}=\boldsymbol{L}\boldsymbol{e}_j$$

$$\begin{aligned}&\mathcal{L}(\boldsymbol{x}_j-\boldsymbol{x}_0)\\&=\mathcal{L}(\boldsymbol{x}_0+t\boldsymbol{e}_j-\boldsymbol{x}_0)\\&=\mathcal{L}(t\boldsymbol{e}_j)\\&=\boldsymbol{L}(t\boldsymbol{e}_j)\\&=t\boldsymbol{L}\boldsymbol{e}_j\end{aligned}$$

The Derivative Matrix

$$\lim_{t\to 0}\frac{\boldsymbol{f}(\boldsymbol{x}_j)-\boldsymbol{f}(\boldsymbol{x}_0)}{t}=\boldsymbol{L}\boldsymbol{e}_j$$

- But $\boldsymbol{L}\boldsymbol{e}_j$ is the *j*th column of the matrix $\boldsymbol{L}$. The vector $\boldsymbol{x}_j$ differs from $\boldsymbol{x}_0$ only in the *j*th coordinate, and in that coordinate the difference is just the number *t*. Therefore, the left side is the partial derivative $\frac{\partial \boldsymbol{f}}{\partial x_j}(\boldsymbol{x}_0)$
- Because vector limits are computed by taking the limit of each coordinate function, it follows that if

$$\boldsymbol{f}(\boldsymbol{x})=\begin{bmatrix} f_1(\boldsymbol{x}) \\ \vdots \\ f_m(\boldsymbol{x}) \end{bmatrix} \qquad \left(\boldsymbol{f}_i : R^n \to R, i=1,...,m\right)$$

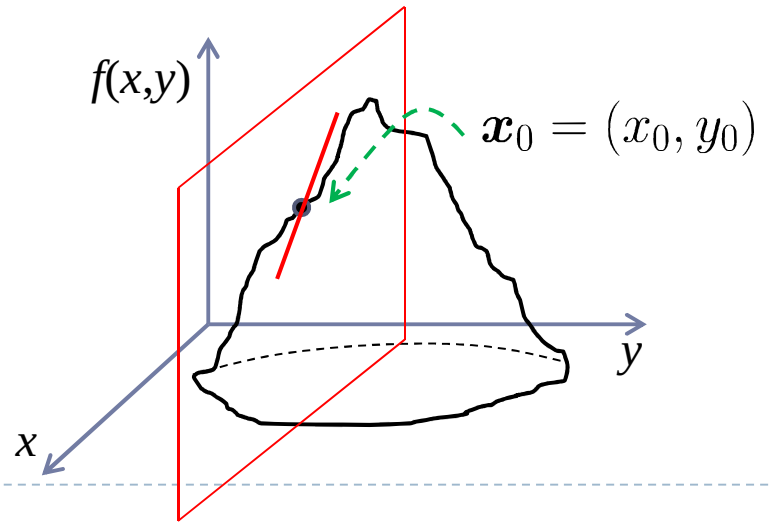
then $\frac{\partial \boldsymbol{f}}{\partial x_j}(\boldsymbol{x}_0)=\begin{bmatrix} \frac{\partial f_1}{\partial x_j}(\boldsymbol{x}_0) \\ \vdots \\ \frac{\partial f_m}{\partial x_j}(\boldsymbol{x}_0) \end{bmatrix}$ and the matrix $\boldsymbol{L}$ has the form

$$\left[\frac{\partial \boldsymbol{f}}{\partial x_1}(\boldsymbol{x}_0)\cdots\frac{\partial \boldsymbol{f}}{\partial x_n}(\boldsymbol{x}_0)\right]=\begin{bmatrix} \frac{\partial f_1}{\partial x_1}(\boldsymbol{x}_0) & \cdots & \frac{\partial f_1}{\partial x_n}(\boldsymbol{x}_0) \\ \vdots & & \vdots \\ \frac{\partial f_m}{\partial x_1}(\boldsymbol{x}_0) & \cdots & \frac{\partial f_m}{\partial x_n}(\boldsymbol{x}_0) \end{bmatrix}$$

The Derivative Matrix

- The matrix $\boldsymbol{L}$ is called the ***Jacobian matrix***, or ***derivative matrix***, of $\boldsymbol{f}$ at $\boldsymbol{x}_0$, and is denoted $D\boldsymbol{f}(\boldsymbol{x}_0)$.
- For convenience, we often refer to $D\boldsymbol{f}(\boldsymbol{x}_0)$ simply as the derivative of $\boldsymbol{f}$ at $\boldsymbol{x}_0$.

$$\left[\frac{\partial \boldsymbol{f}}{\partial x_1}(\boldsymbol{x}_0) \cdots \frac{\partial \boldsymbol{f}}{\partial x_n}(\boldsymbol{x}_0)\right] = \begin{bmatrix} \frac{\partial f_1}{\partial x_1}(\boldsymbol{x}_0) & \cdots & \frac{\partial f_1}{\partial x_n}(\boldsymbol{x}_0) \\ \vdots & & \vdots \\ \frac{\partial f_m}{\partial x_1}(\boldsymbol{x}_0) & \cdots & \frac{\partial f_m}{\partial x_n}(\boldsymbol{x}_0) \end{bmatrix}$$



$$\mathcal{A}(x) = f(x_0) + f'(x_0)(x - x_0)$$

The Derivative Matrix

- Theorem 5.5: If a function $\boldsymbol{f} : R^n \to R^m$ is differentiable at $\boldsymbol{x}_0$, then the derivative of $\boldsymbol{f}$ at $\boldsymbol{x}_0$ is determined uniquely and is represented by an $m \times n$ derivative matrix $D\boldsymbol{f}(\boldsymbol{x}_0)$. The best affine approximation of $\boldsymbol{f}$ near $\boldsymbol{x}_0$ is then given by

$$\mathcal{A}(\boldsymbol{x}) = \boldsymbol{f}(\boldsymbol{x}_0) + D\boldsymbol{f}(\boldsymbol{x}_0)(\boldsymbol{x} - \boldsymbol{x}_0)$$

in the sense that

$$\boldsymbol{f}(\boldsymbol{x}) = \mathcal{A}(\boldsymbol{x}) + \boldsymbol{r}(\boldsymbol{x})$$

and $\lim_{\boldsymbol{x} \to \boldsymbol{x}_0} \|\boldsymbol{r}(\boldsymbol{x})\| / \|\boldsymbol{x} - \boldsymbol{x}_0\| = 0$.

The columns of the derivative matrix $D\boldsymbol{f}(\boldsymbol{x}_0)$ are vector partial derivatives. The vector $\frac{\partial \boldsymbol{f}}{\partial x_j}(\boldsymbol{x}_0)$ is a tangent vector at $\boldsymbol{x}_0$ to the curve $\boldsymbol{f}$ obtained by varying only the *j*th coordinate of $\boldsymbol{x}$.

The Derivative Matrix

- If $f : R^n \to R$ is differentiable, then the function $\bigtriangledown f$ defined by

$$\bigtriangledown f(\boldsymbol{x}) = \begin{bmatrix} \frac{\partial f}{\partial x_1}(\boldsymbol{x}) \\ \vdots \\ \frac{\partial f}{\partial x_n}(\boldsymbol{x}) \end{bmatrix} = Df(\boldsymbol{x})^T$$

  is called the ***gradient*** of f .

- Given $f : R^n \to R$, if $\bigtriangledown f$ is differentiable, we say that f is ***twice differentiable***, and we write the derivative of $\bigtriangledown f$ as

$$D^2 f = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_2 \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_1} \\ \frac{\partial^2 f}{\partial x_1 \partial x_2} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_2} \\ \vdots & & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_1 \partial x_n} & \frac{\partial^2 f}{\partial x_2 \partial x_n} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix}$$

  $\frac{\partial^2 f}{\partial x_i \partial x_j}$ represents taking the partial derivative with respect to x_j first, then with respect to x_i

The Derivative Matrix

- The matrix $D^2 f(\boldsymbol{x})$ is called the ***Hessian matrix*** of f at $\boldsymbol{x}$, and is often also denoted $\boldsymbol{F}(\boldsymbol{x})$.
- A function $\boldsymbol{f} : \Omega \to R^m, \Omega \subset R^n$, is said to be ***continuously differentiable*** on Ω if it is differentiable (on Ω), and $D\boldsymbol{f} : \Omega \to R^{m \times n}$ is continuous; that is, the components of $\boldsymbol{f}$ have continuous partial derivatives. In this case, we write $\boldsymbol{f} \in \mathcal{C}^1$. If the components of $\boldsymbol{f}$ have continuous partial derivatives of order p, we write $\boldsymbol{f} \in \mathcal{C}^p$.
- The Hessian matrix of a function $f : R^n \to R$ at $\boldsymbol{x}$ is symmetric if f is twice continuously differentiable at $\boldsymbol{x}$. This is a well-known result from calculus called ***Clairaut's theorem*** or ***Schwarz's theorem***.

Example

$$\frac{d}{dx}\left(\frac{f}{g}\right) = \frac{f'g - fg'}{g^2}$$

$$\boldsymbol{F} = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_2 \partial x_1} \\ \frac{\partial^2 f}{\partial x_1 \partial x_2} & \frac{\partial^2 f}{\partial x_2^2} \end{bmatrix}$$

- Consider the function $f(\boldsymbol{x}) = \begin{cases} x_1 x_2 (x_1^2 - x_2^2)/(x_1^2 + x_2^2) & \text{if } \boldsymbol{x} \neq \mathbf{0} \\ 0 & \text{if } \boldsymbol{x} = \mathbf{0} \end{cases}$

  Compute its Hessian at the point $\mathbf{0} = [0, 0]^T$

- Start with $\dfrac{\partial^2 f}{\partial x_1^2} = \dfrac{\partial}{\partial x_1}\left(\dfrac{\partial f}{\partial x_1}\right)$

  $$\frac{\partial f}{\partial x_1}(\boldsymbol{x}) = \begin{cases} x_2 (x_1^4 - x_2^4 + 4x_1^2 x_2^2)/(x_1^2 + x_2^2)^2 & \text{if } \boldsymbol{x} \neq \mathbf{0} \\ 0 & \text{if } \boldsymbol{x} = \mathbf{0} \end{cases}$$

  Note that

  $$\frac{\partial f}{\partial x_1}([x_1, 0]^T) = 0 \qquad \frac{\partial^2 f}{\partial x_1^2}(\mathbf{0}) = 0$$

  $$\frac{\partial f}{\partial x_1}([0, x_2]^T) = -x_2 \qquad \frac{\partial f}{\partial x_2 x_1}(\mathbf{0}) = -1$$

Example

- We next compute

$$\frac{\partial^2 f}{\partial x_2^2} = \frac{\partial}{\partial x_2}\left(\frac{\partial f}{\partial x_2}\right)$$

$$\frac{\partial f}{\partial x_2}(\boldsymbol{x}) = \begin{cases} x_1(x_1^4 - x_2^4 - 4x_1^2x_2^2)/(x_1^2 + x_2^2)^2 & \text{if } \boldsymbol{x} \neq \mathbf{0} \\ 0 & \text{if } \boldsymbol{x} = \mathbf{0} \end{cases}$$

$$\frac{\partial f}{\partial x_2}([0, x_2]^T) = 0 \qquad \frac{\partial f}{\partial x_2^2}(\mathbf{0}) = 0$$

$$\frac{\partial f}{\partial x_2}([x_1, 0]^T) = x_1 \qquad \frac{\partial f}{\partial x_1 x_2}([x_1, 0]^T) = 1$$

- Therefore, the Hessian evaluated at the point $\mathbf{0} = [0, 0]^T$ is

$$\boldsymbol{F}(\mathbf{0}) = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$$

Differentiation Rules

$$\frac{d}{dx}g(f(x)) = g'(f(x))f'(x)$$

- The ***chain rule*** for differentiating the composition $g(\boldsymbol{f}(t))$, of a function $\boldsymbol{f} : R \to R^n$ and a function $g : R^n \to R$
- Theorem 5.6: Let $g : \mathcal{D} \to R$ be differentiable on an open set $\mathcal{D} \subset R^n$, and let $f : (a, b) \to \mathcal{D}$ be differentiable on (a, b). Then, the composite function $h : (a, b) \to R$ given by $h(t) = g(\boldsymbol{f}(t))$ is differentiable on (a, b), and

$$h'(t) = Dg(\boldsymbol{f}(t))D\boldsymbol{f}(t) = \bigtriangledown g(\boldsymbol{f}(t))^T \begin{bmatrix} f_1'(t) \\ \vdots \\ f_n'(t) \end{bmatrix}$$

Differentiation Rules

$$\frac{d}{dx}(fg) = f'g + fg'$$

- ***Product rule***: Let $\boldsymbol{f} : R^n \to R^m$ and $\boldsymbol{g} : R^n \to R^m$ be two differentiable functions. Define the function $h : R^n \to R$ by $h(x) = \boldsymbol{f}(\boldsymbol{x})^T \boldsymbol{g}(\boldsymbol{x})$. Then h is also differentiable and

$$Dh(x) = \boldsymbol{f}(\boldsymbol{x})^T D\boldsymbol{g}(\boldsymbol{x}) + \boldsymbol{g}(\boldsymbol{x})^T D\boldsymbol{f}(\boldsymbol{x})$$

- Some useful formulas: Let $\boldsymbol{A} \in R^{m \times n}$ and $\boldsymbol{y} \in R^m$, the derivative with respect to $\boldsymbol{x}$

$$D(\boldsymbol{y}^T \boldsymbol{A}\boldsymbol{x}) = \boldsymbol{y}^T \boldsymbol{A}$$

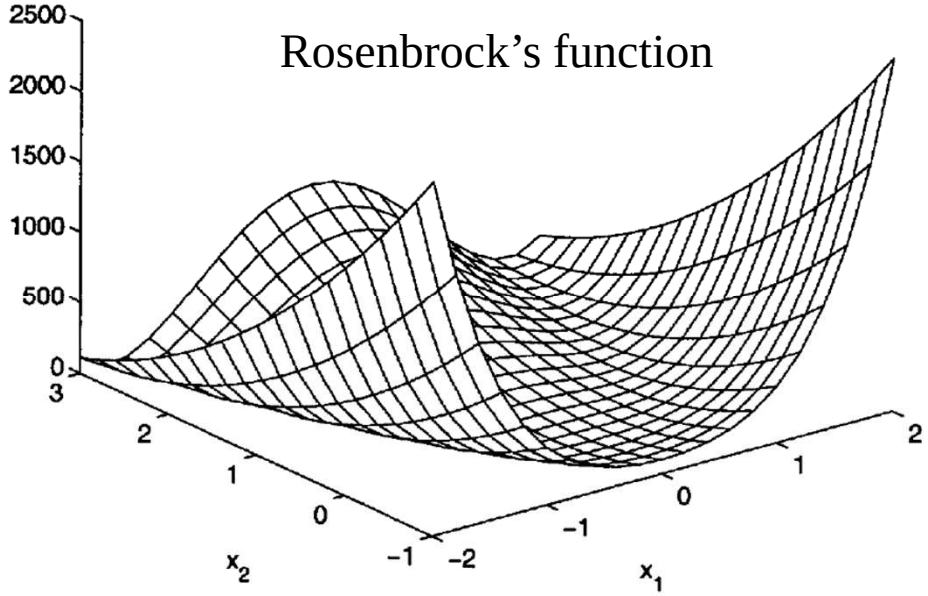
$$D(\boldsymbol{x}^T \boldsymbol{A}\boldsymbol{x}) = \boldsymbol{x}^T(\boldsymbol{A} + \boldsymbol{A}^T), \quad \text{if } m = n$$

- If $\boldsymbol{y} \in R^n$, then $D(\boldsymbol{y}^T \boldsymbol{x}) = \boldsymbol{y}^T$
- If $\boldsymbol{Q}$ is a symmetric matrix, then $D(\boldsymbol{x}^T \boldsymbol{Q}\boldsymbol{x}) = 2\boldsymbol{x}^T \boldsymbol{Q}$. In particular, $D(\boldsymbol{x}^T \boldsymbol{x}) = 2\boldsymbol{x}^T$

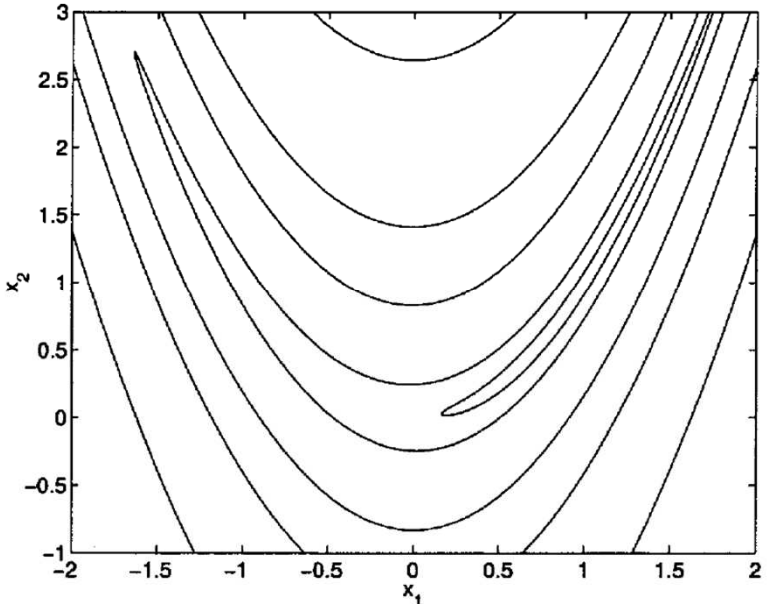
Level Sets and Gradients

- The ***level set*** of a function $f : R^n \rightarrow R$ at level c is the set of points $S = \{\boldsymbol{x} : f(\boldsymbol{x}) = c\}$. For $f : R^2 \rightarrow R$, we are usually interested in S when it is a curve. For $f : R^3 \rightarrow R$, the sets S most often considered are surfaces.
- Example: Consider $f(\boldsymbol{x}) = 100(x_2 - x_1^2)^2 + (1 - x_1)^2, \boldsymbol{x} = [x_1, x_2]^T$
  It is called ***Rosenbrock's function***.

Rosenbrock's function



Level sets at levels 0.7, 7, 70, 200, and 700



Level Sets and Gradients

- A point $\boldsymbol{x}_0$ is on the level set S at level c means that $f(\boldsymbol{x}_0) = c$. Suppose that there is a curve γ lying in S and parameterized by a continuously differentiable function $\boldsymbol{g} : R \to R^n$. Suppose also that $\boldsymbol{g}(t_0) = \boldsymbol{x}_0$ and $D\boldsymbol{g}(t_0) = \boldsymbol{v} \neq 0$, so that $\boldsymbol{v}$ is a tangent vector to γ at $\boldsymbol{x}_0$.
- Applying the chain rule to the function $h(t) = f(\boldsymbol{g}(t))$

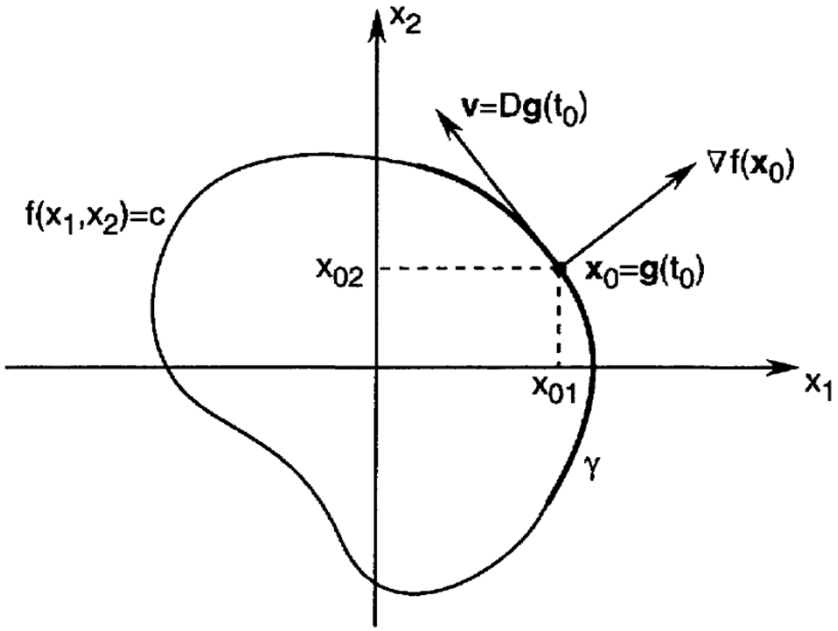
  $h'(t_0) = Df(\boldsymbol{g}(t_0))D\boldsymbol{g}(t_0) = Df(\boldsymbol{x}_0)\boldsymbol{v}$

  Since γ lies on S, we have

  $h(t) = f(\boldsymbol{g}(t)) = c$

  That is, h is constant. Thus,

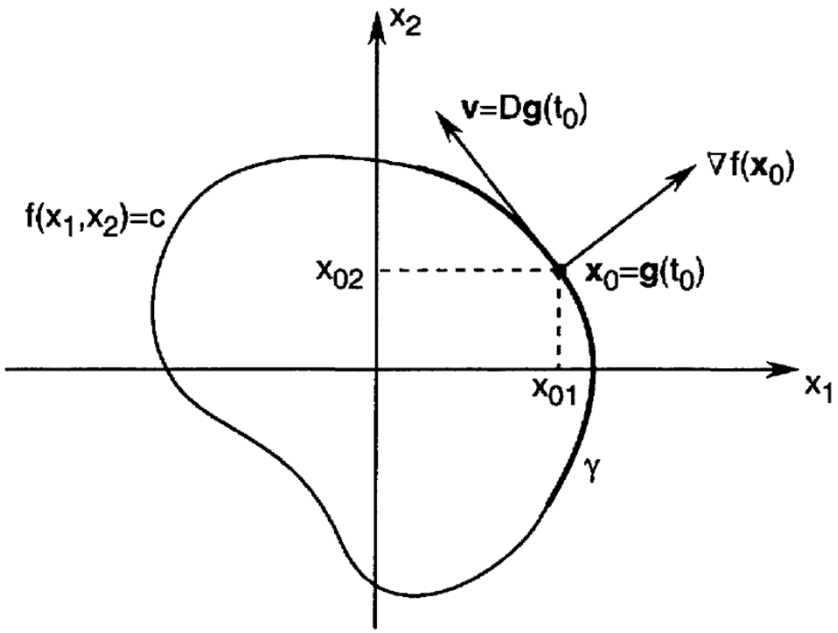
  $h'(t_0) = 0 = Df(x_0)\boldsymbol{v} = \bigtriangledown f(\boldsymbol{x}_0)^T\boldsymbol{v}$



Level Sets and Gradients

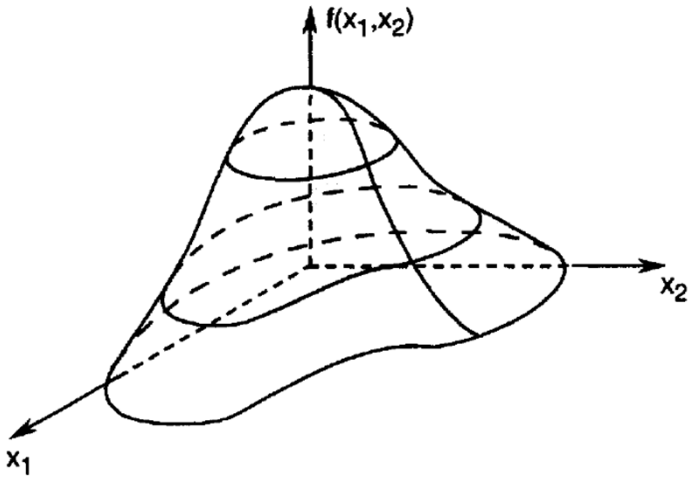
- Theorem 5.7: The vector $\bigtriangledown f(\boldsymbol{x}_0)$ is orthogonal to the tangent vector to an arbitrary smooth curve passing through $\boldsymbol{x}_0$ on the level set determined by $f(\boldsymbol{x}) = f(\boldsymbol{x}_0)$
- It is natural to say that $\bigtriangledown f(\boldsymbol{x}_0)$ is ***orthogonal*** or ***normal*** to the level set S corresponding to $\boldsymbol{x}_0$, and to take as the tangent plane (or line) to S at $\boldsymbol{x}_0$ the set of all points $\boldsymbol{x}$ satisfying

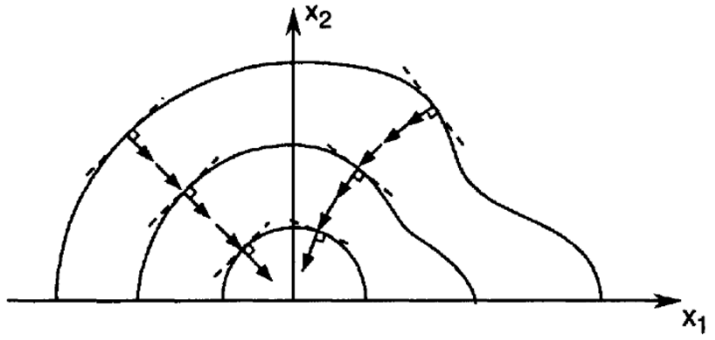
$$\bigtriangledown f(\boldsymbol{x}_0)^T(\boldsymbol{x} - \boldsymbol{x}_0) = 0, \quad \text{if } \bigtriangledown f(\boldsymbol{x}_0) \neq \mathbf{0}$$



Level Sets and Gradients

- $\bigtriangledown f(\boldsymbol{x}_0)$ is the direction of ***maximum rate of increase*** of f at $\boldsymbol{x}_0$
- The direction of maximum rate of increase of a real-valued differentiable function at a point is orthogonal to the level set of the function through that point.
- An example about $f : R^2 \rightarrow R$
  - The curve on the surface running from bottom to top has the property that its projection onto the (x_1, x_2)-plane is always orthogonal to the level curves and is called a ***path of steepest ascent***.





Level Sets and Gradients

- The ***graph*** of $f : R^n \to R$ is the set $\{[\boldsymbol{x}^T, f(\boldsymbol{x})]^T : \boldsymbol{x} \in R^n\} \subset R^{n+1}$. The notion of the gradient of a function has an alternative useful interpretation in terms of the tangent hyperplane to its graph.
- Let $\boldsymbol{x}_0 \in R^n$ and $z_0 = f(\boldsymbol{x}_0)$. The point $[\boldsymbol{x}_0^T, z_0]^T \in R^{n+1}$ is a point on the graph of f. If f is differentiable at $\boldsymbol{\xi}$, then the graph admits a nonvertical tangent hyperplane at $\boldsymbol{\xi} = [\boldsymbol{x}^T, z_0]^T$. The hyperplane through $\boldsymbol{\xi}$ is the set of all points $[x_1, ..., x_n, z]^T \in R^{n+1}$ satisfying the equation

$$u_1(x_1 - x_{01}) + \cdots + u_n(x_n - x_{0n}) + v(z - z_0) = 0$$

where the vector $[u_1, ..., u_n, v] \in R^{n+1}$ is normal to the hyperplane.

$$\langle (u_1, ..., u_n, v), (x_1 - x_{01}, ..., x_n - x_{0n}, z - z_0) \rangle \qquad \boldsymbol{x}_0 = (x_{01}, ..., x_{0n})$$

normal — a vector on the hyperplane

Level Sets and Gradients

$u_1(x_1 - x_{01}) + \cdots + u_n(x_n - x_{0n}) + v(z - z_0) = 0$

- Assuming that this hyperplane is nonvertical (that is, $v \neq 0$), let

$$d_i = -\frac{u_i}{v}$$

Thus, we can rewrite the hyperplane equation as

$$z = d_1(x_1 - x_{01}) + \cdots + d_n(x_n - x_{0n}) + z_0$$

- We can think of the right side as a function $z : R^n \to R$. Observe that for the hyperplane to be tangent to the graph of f, the functions f and z must have the same partial derivatives at the point $\boldsymbol{x}_0$. Hence, if f is differential at $\boldsymbol{x}_0$, its tangent hyperplane can be written in terms of its gradient, as given by the equation

$$z - z_0 = Df(\boldsymbol{x}_0)(\boldsymbol{x} - \boldsymbol{x}_0) = (\boldsymbol{x} - \boldsymbol{x}_0)^T \bigtriangledown f(\boldsymbol{x}_0)$$

$$Df(\boldsymbol{x}_0) = \frac{z - z_0}{\boldsymbol{x} - \boldsymbol{x}_0}$$

Taylor Series

- Theorem 5.8 ***Taylor's Theorem***: Assume that a function $f : R \to R$ is m times continuously differentiable (i.e. $f \in \mathcal{C}^m$) on an interval $[a, b]$. Denote $h = b - a$

  Then,

  $$f(b) = f(a) + \frac{h}{1!}f^{(1)}(a) + \frac{h^2}{2!}f^{(2)}(a) + \cdots + \frac{h^{m-1}}{(m-1)!}f^{(m-1)}(a) + R_m$$

  (called Taylor's formula) where $f^{(i)}$ is the *i*th derivative of f, and

  $$R_m = \frac{h^m(1-\theta)^{m-1}}{(m-1)!}f^{(m)}(a + \theta h) = \frac{h^m}{m!}f^{(m)}(a + \theta' h) \qquad \theta, \theta' \in (0, 1)$$

Taylor Series

- An important property of Taylor's theorem arises from the forms of the remainder R_m.
- We introduce the ***order symbols***, O and o.
- Let g be a real-valued function defined in some neighborhood of $\mathbf{0} \in R^n$, with $g(\boldsymbol{x}) \neq 0$ if $\boldsymbol{x} \neq \mathbf{0}$. Let $\boldsymbol{f} : \Omega \to R^m$ be defined in a domain $\Omega \subset R^n$ that includes $\mathbf{0}$. Then, we write
  - 1. $\boldsymbol{f}(\boldsymbol{x}) = O(g(\boldsymbol{x}))$ to mean that the quotient $\|\boldsymbol{f}(\boldsymbol{x})\|/|g(\boldsymbol{x})$ is bounded near $\mathbf{0}$; that is, these exist numbers $K > 0$ and $\delta > 0$ such that if $\|\boldsymbol{x}\| < \delta$, $\boldsymbol{x} \in \Omega$, then $\|\boldsymbol{f}(\boldsymbol{x})\|/|g(\boldsymbol{x})| \leq K$ or $\|\boldsymbol{f}(\boldsymbol{x})\| \leq K g(\boldsymbol{x})$
  - 2. $\boldsymbol{f}(\boldsymbol{x}) = o(g(\boldsymbol{x}))$ to mean that

$$\lim_{\boldsymbol{x} \to \mathbf{0}, \boldsymbol{x} \in \Omega} \frac{\|\boldsymbol{f}(\boldsymbol{x})\|}{|g(\boldsymbol{x})|} = 0$$

Taylor Series

- The symbol $O(g(x))$ [read "big-oh" of $g(x)$] is used to represent a function that is bounded by a scaled version of g in a neighborhood of 0.
- Examples:
  - $x = O(x)$
  - $\begin{bmatrix} x^3 \\ 2x^2 + 3x^4 \end{bmatrix} = O(x^2)$
  - $\cos x = O(1)$
  - $\sin x = O(x)$

Taylor Series

- On the other hand, $o(g(\boldsymbol{x}))$ [read "little-oh" of $g(\boldsymbol{x})$] represents a function that goes to zero "faster" than $g(\boldsymbol{x})$ in the sense that $\lim_{\boldsymbol{x}\to\mathbf{0},\boldsymbol{x}\in\Omega}\|o(g(\boldsymbol{x}))\|/|g(\boldsymbol{x})| = 0$
- Examples:
  - $x^2 = o(x)$
  - $\begin{bmatrix} x^3 \\ 2x^2 + 3x^4 \end{bmatrix} = o(x)$
  - $x^3 = o(x^2)$
  - $x = o(1)$

Taylor Series

- Note that if $\boldsymbol{f}(\boldsymbol{x}) = o(g(\boldsymbol{x}))$, then $\boldsymbol{f}(\boldsymbol{x}) = O(g(\boldsymbol{x}))$ (but the converse is not necessarily true). Also, if $\boldsymbol{f}(\boldsymbol{x}) = O(\|\boldsymbol{x}\|^p)$, then $\boldsymbol{f}(\boldsymbol{x}) = o(\|\boldsymbol{x}\|^{p-\epsilon})$ for any $\epsilon > 0$
- Suppose that $f \in \mathcal{C}^m$. Recall that the remainder term in Taylor's theorem has the form

$$R_m = \frac{h^m}{m!} f^{(m)}(a + \theta h) \qquad \theta \in (0, 1)$$

Substituting this into Taylor's formula, we get

$$f(b) = f(a) + \frac{h}{1!} f^{(1)}(a) + \frac{h^2}{2!} f^{(2)}(a) + \cdots + \frac{h^{m-1}}{(m-1)!} f^{(m-1)}(a) + \frac{h^m}{m!} f^{(m)}(a + \theta h)$$

By the continuity of $f^{(m)}$, we have $f^{(m)}(a + \theta h) \to f^{(m)}(a)$ as $h \to 0$ that is, $f^{(m)}(a + \theta h) = f^{(m)}(a) + o(1)$. Therefore,

$$R_m = \frac{h^m}{m!} f^{(m)}(a + \theta h) = \frac{h^m}{m!} f^{(m)}(a) + o(h^m) \qquad h^m o(1) = o(h^m)$$

Taylor Series

- We may then write Taylor's formula as

$$f(b) = f(a) + \frac{h}{1!}f^{(1)}(a) + \frac{h^2}{2!}f^{(2)}(a) + \cdots + \frac{h^{m-1}}{(m-1)!}f^{(m-1)}(a) + \frac{h^m}{m!}f^{(m)}(a) + o(h^m)$$

- If, in addition, we assume that $f \in \mathcal{C}^{m+1}$, we may replace the term $o(h^m)$ above by $O(h^{m+1})$

$$f(b) = f(a) + \frac{h}{1!}f^{(1)}(a) + \frac{h^2}{2!}f^{(2)}(a) + \cdots + \frac{h^{m-1}}{(m-1)!}f^{(m-1)}(a) + \frac{h^m}{m!}f^{(m)}(a) + O(h^{m+1})$$

Taylor Series

- Theorem 5.9 ***Mean value theorem***: If a function $\boldsymbol{f} : R^n \to R^m$ is differentiable on an open set $\Omega \subset R^n$, then for any pair of points $\boldsymbol{x}, \boldsymbol{y} \in \Omega$, there exists a matrix $\boldsymbol{M}$ such that

$$\boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{f}(\boldsymbol{y}) = \boldsymbol{M}(\boldsymbol{x} - \boldsymbol{y})$$

- The mean value theorem follows from Taylor's theorem (for the case where $m = 1$) applied to each component of $\boldsymbol{f}$. $\boldsymbol{M}$ is a matrix whose rows are the rows of $D\boldsymbol{f}$ evaluated at points that lie one the line segment joining $\boldsymbol{x}$ and $\boldsymbol{y}$.

若函數f(x)在[a,b]區間上連續並可維分，
則在該區間內必存在一點c，使

$$f'(c) = \frac{f(b)-f(a)}{b-a}$$